

Developing a Scientific Workforce Analysis and Modeling Framework (SWAM)

Michael Larsen (GWU)

With Siyu Qing, Beilei Zhou, Mary Foulkes

Joshua Hawley (Ohio State), Richard Larson (MIT)

With Emrah Cimren, Joseph Fiksel, Anand Desai

Overview

The Scientific Workforce Analysis and Modeling project (SWAM), sponsored by NIGMS, focuses on modeling the dynamics of the scientific workforce to help policy makers evaluate alternatives and anticipate the consequences of their decisions. Two projects were funded by SWAM in the initial competition.

1) Researchers at The George Washington University (GWU) are developing methods for longitudinal analysis of NSF Survey of Doctorate Recipient data and estimating cross-sectional characteristics of postdoctoral researchers using NSF Survey of Doctorate Recipient data.

2) The Ohio State University's Battelle Center and MIT faculty are developing a model to evaluate the labor market transitions of graduates from Ph.D. programs in biomedical sciences.

The document below gives some common background and then describes the survey data analysis of the GWU team and the agent-based modeling of the OSU-MIT team.

Background

In the U.S., women and minorities are severely underrepresented in many academic and nonacademic career areas of medicine & health (M&H) and science & engineering (S&E). They also are underrepresented in NIH grant applications and awards. This is a critical issue for maintaining the size and productivity of these workforces in the U.S. The principal goal of the research by the two teams is to use existing sources of information in novel ways to address key questions about M&H and S&E workforce.

One particular question of interest is the following: what factors influence who becomes a NIH-funded researcher? At least four concepts appear in the literature:

1. Pipeline.

2. Life course processes.
3. Key transitions.
4. Model flows or dynamics, multi-state life table.

An example of the pipeline is high school-college-graduate school-postdoctoral position-assistant professor-associate professor-full professor. Focusing on the pipeline analogy leads one to ask, where are the leakages of individuals out of the research career track? Life course processes expand the idea of a pipeline to acknowledge that several paths are possible, including ones involving time out of the workforce, part-time work, and returns to a research career track. This conceptualization is particularly important when considering different experiences of men and women and individuals with and without children. Instead of examining the long-term career track, one can instead focus on key transitions, such as earning a Ph.D., getting a postdoctoral position, or receiving tenure. Model flows or dynamics and multi-state life tables examine transitions at an aggregate level. As more factors are included in the dynamic flows, these approaches can resemble the micro-data modeling employed in the other approaches.

To study these concepts, one needs data over time. Cross sectional data concern a population at a given time frame. Cross sectional data have been used to examine progress toward achievements, such as getting a Ph.D. or receiving tenure, for a cohort of individuals. One could repeat such an analysis for different cohorts over time in order to examine the stability of relationships among variables over time. Cross sectional data are limited in their ability, however, to describe change. One cannot, for example, estimate career path probabilities for individuals with cross sectional data alone. Longitudinal data record information over time for a given set of individuals. In the present context, one could use such data to describe actual career paths over time. The NSF Survey of Doctoral Recipients (SDR) data have some elements of both cross sectional and longitudinal studies, as described below.

Additionally, modeling (e.g., Agent Based Modeling or other techniques) has been used to understand the interplay between individual choices, institutional policies, and larger social and demographic factors. Although there are no comprehensive models (ABM or otherwise) designed for studying the full biomedical workforce, recent modeling efforts from the STEM Research and Modeling Network (SRMN) have produced baseline models for understanding the factors impacting the STEM workforce. However, these prior efforts have used limited data and focused on the high school to college transition. Therefore, there is a critical need to build a modeling framework specific to the biomedical workforce. The *objective* of this project is to integrate existing knowledge into an innovative, flexible modeling framework and to demonstrate the applicability of this framework.

Data

The research projects funded under SWAM take advantage of sophisticated longitudinal data collected the National Science Foundation as well as exploiting existing cross sectional data on the scientific workforce. The data used help understand the influence of individual and institutional factors on the long term education and career outcomes of individuals engaged in the science workforce.

Longitudinal Data

The primary data source being used by the teams at GWU and OSU/MIT is the National Science Foundation's (NSF's) Survey of Doctorate Recipients (SDR). This is a cross sectional survey conducted every 2 or 3 years. Each survey year, the target population is a little bit different because people enter (e.g., new PhD recipients in the U.S.) or leave (e.g., deaths) the population. Survey weights adjust for oversampling and nonresponse on a cross sectional basis. Thus, it is important to use survey weights in order to have unbiased estimation for a population total in a given survey year.

Variables included in the data set include

- Labor force status
- Source of funding
- Academic rank and tenure
- Salary
- Field, institution of degree, employment
- Demographics: age, sex, race/ethnicity, marital, spouse work, child at home, child age, U.S. citizenship
- Work responsibilities, management position
- Professional memberships
- Reasons for taking postdoctoral position
- Questions about a career path job.

A significant portion of the sample (e.g., 60% on 3 or more surveys from 1993-2006) appears in multiple survey years and can be linked across time. No longitudinal weight exists to enable estimation of statistical models or comparison of finite population characteristics using data from multiple survey waves together. Instead, one can take a single survey year and estimate career paths for individuals from that year.

The advantage of combining data from survey years is an increase in sample size versus a single cohort. Although the NSF SDR survey is large by most standards, the number of individuals in certain discipline by rank by demographic group combinations in a single survey year can be small. One complication with combining data from different survey years is that each individual in each year has survey weight for that year. For an individual with multiple weights, which weight should be used? This is an open research question in survey sampling statistical theory and practice. There are two other

complications. Some individuals, such as recent Ph.D. recipients, are not members of the population until they obtain their Ph.D. Analyses might logically exclude some individuals for some relationships due to this fact. A more serious complication is variance estimation. The SDR estimates variances using a technique that is specific to each year, so any longitudinal weight will need to determine how to address the different variance components.

Cross Sectional Data

Additionally, the OSU team is making use of cross sectional data from the AAAS that is collected on an annual basis from both postdoctoral recipients and advisors of postdoctoral recipients. Science Careers Journal¹, which is produced by AAAS, conducts an annual survey to gather data starting from 2004. The survey alternates each year starting from 2004 between asking the opinions of postdocs and postdoc supervisors. The survey includes over 8800 postdoctoral recipients from 2004-2010 and 1778 postdoctoral supervisors from 2005-2009.

The survey data from Science Careers on postdoctoral recipients are analyzed to determine what factors influence postdoctoral recipients to select research careers in medicine & health (M&H). The simulation models also investigate criteria of supervisors in recruiting postdocs. The data from the AAAS analysis are used to identify parameters and decision rules in the ABM simulation. While these data lack the longitudinal nature that the NSF data have, they contain critical variables that will be useful to developing models of the decision making process of postdoctoral associates. The variables include the following:

- Job search data (e.g., job availability, individual search techniques, career goals)
- Employment positions (e.g., postdoctoral and faculty positions) in both academic and industry.
- Evaluating applicants and making offers (e.g., information on the process by which organizations assess postdoctoral applicants and weigh offers)

Work to Date

George Washington University

Creation of longitudinal weights using survey weight calibration

Some surveys are designed to enable estimation over time using longitudinal data. The National Resources Inventory of the USDA, for example, has as its sample frame the land area of the United States. An initial selection of land area was made using a probability sample. Subsequent yearly samples are taken from the first stage sample with probabilities based on land cover and land use in the foundational sample. Yearly

¹ <http://sciencecareers.sciencemag.org/>

samples are split into a panel on which data are gathered every year and rotations that then are sampled in subsequent years. Probabilities across time then are consistent.

The American Community Survey of the U.S. Census Bureau has as its sample frame all the addresses in the United States. Samples are selected for five years a time. This simultaneous selection enables the development of consistent weights over time and direct aggregation of results for 3- and 5-year estimates at lower levels of geography.

The core task of the GWU team is to develop new weights for the SDR that will allow for the longitudinal analysis of the doctoral recipients over time. To carry out this task the researchers will generate new survey and sampling weights using a calibration method. This method will require that the researchers take into account the weights that were generated from each year data were collected and develop new weights that can be used across the different data files.

The new single weight that is generated will meet three requirements. First, the weight needs to be calculable from existing data, which means either the public use data sets or the restricted use versions that NSF releases under strict licensing. Second, the weight needs to be useful for reproducing key cross sectional analyses. Third, the weight should be low in variability, because high variability weights are associated with low precision in estimation

The calibration ideas were applied to a few variables for three years (1993, 1995, and 1997) from the NSF SDR public use data files. This was done before the team had received the restricted-use data. Initial evidence suggests that calibration can create useful longitudinal weights. Weights preserve means and group sizes by year without inflating standard errors. This should enable longitudinal analysis. Further simulation study of this idea is underway.

Postdoctoral researchers (postdocs)

The overall group decided in a previous meeting to study postdoctoral researchers due to their importance to the biomedical and health research workforce. Using the NSF SDR, the GWU team is looking at questions of who gets postdocs and where do postdocs go? In particular, they are examining differences by field, institution, sex, race/ethnicity, marital status, employment status of spouse, and having children at home, especially young children. Tables similar to the examples in Table 1 illustrate the kind of data that have been created. Differences across and trends by major discipline and by demographic groups will also be examined. Ultimately, the team will include data from pre-1993 through 2008.

Table 1: Illustrative Table from SDR Data

Variable	1993			1995			...	2003		
	n	Est	SE	n	Est	SE	...	n	Est	SE
Number of postdocs										
% female										
% Hispanic										
% by race group										
% by US citizenship										
% by marital status										
% with working spouse (if married)										
% with spouse in technical area (if married and spouse working										
% with children at home										
% with children at home under age 6										
% attending a professional meeting										
% attending work related training										
% supported by a grant										
% reporting a										

The data approaches presented here are fundamental to estimating parameters for the modeling work being conducted by the OSU/MIT team.

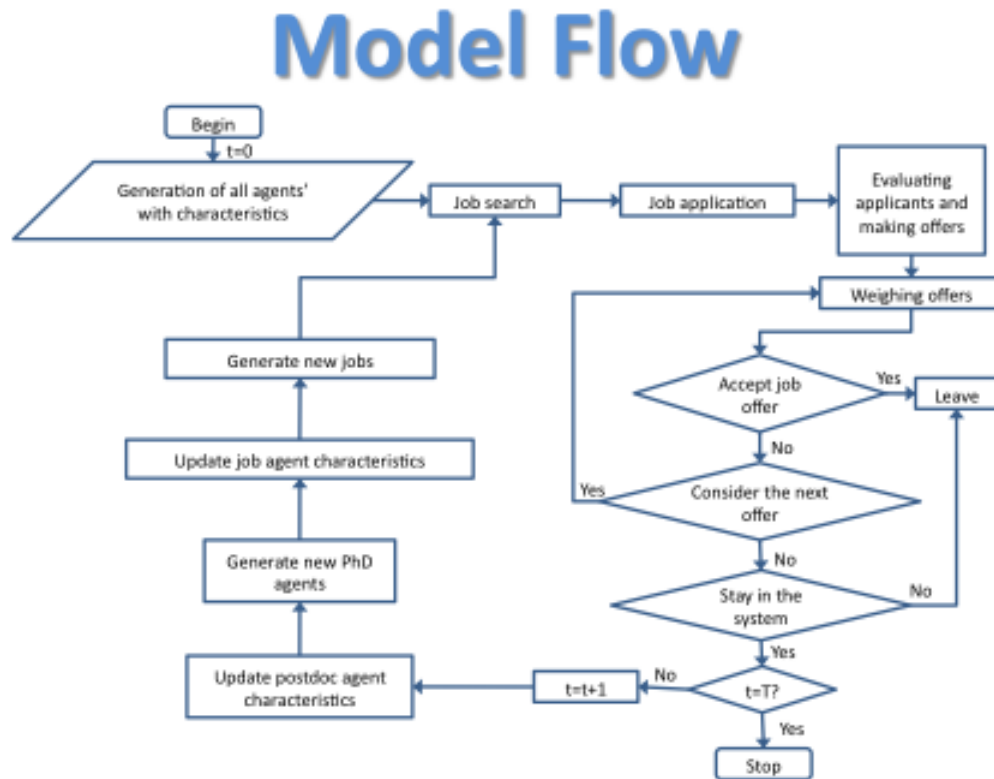
Ohio State University/MIT Agent Based Modeling Framework

Figure 1 illustrates the model that matches graduates with PhD degree (PhD) and postdocs with job providers using employment search methods at the individual level. The outcomes depend on travel preferences of the job searcher, skills, pay, the locations of employment opportunities, and the willingness of companies/institutions to employ the searcher. The geographic area of model contains both employment locations that may be flexibly arranged into one or multiple employment zones or be randomly distributed as well as housing locations. Companies and housing locations are assumed to be exogenous.

The agents in this model are the employment positions, PhDs, and postdocs that interact with one another in determining job opportunities and pay scales, and negotiate agreeable arrangements for employment. Each of these agents is discussed below. Figure 1 describes the model and outlines the decision framework used in the ABM model.

Figure 1 describes a system that conceptualizes the job search characteristics of students that graduate with a degree in bio-medical fields. The model matches graduates in bio-medical fields after graduation with potential jobs using knowledge about the characteristics of graduates and jobs derived from existing surveys collected by the AAAS.

Figure 1: Model Concept Map



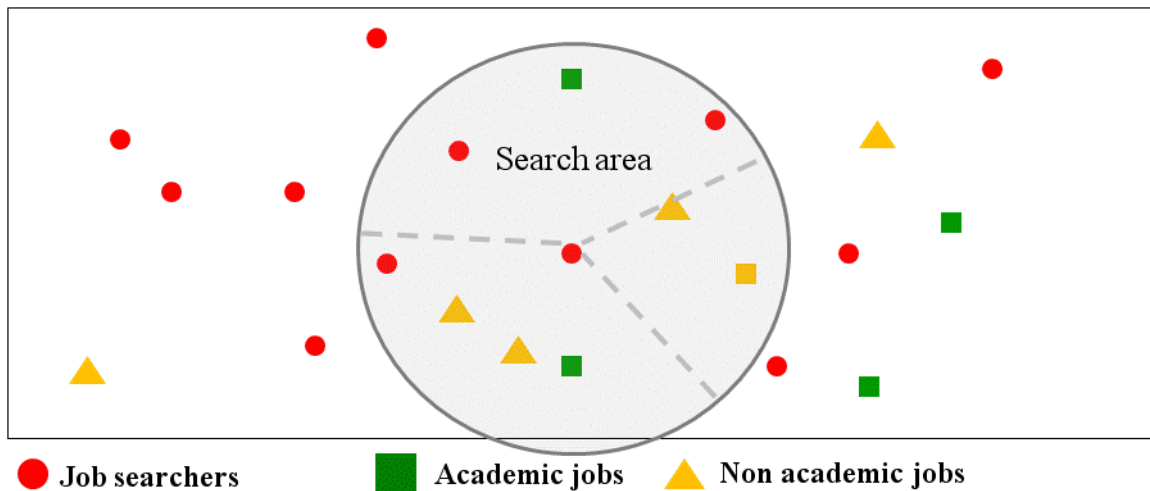
Specifically this model depends on an understanding of the following parameters:

- Characteristics of entering biomedical graduates (age, gender, marital status, children, income, knowledge and skills).
- Employment positions including academic (postdoc and faculty positions) and non-academic (industry and government positions)

- Job search and matching (e.g., labor market data on job availability, individual attributes governing job search intensity, career goals).

Figure 2 provides a way of understanding the ABM model and how it operationalizes the job search techniques. Individual job searchers are matched with either academic or non-academic jobs (e.g., industry) in a specific geographical zone called a search area. This is denoted in Figure 2 by the circle.

Figure 2: Illustrating the Job Search²



- Evaluating applicants and making offers (e.g., information on the process by which organizations assess postdoctoral applicants and weigh offers). The key to this particular iteration in the model is to see if applicants' characteristics (see above) are a match with the skill requirements of the jobs themselves. Figure 3 below provides a graphical representation of that process wherein applicants and jobs interact.
- Weighing offers, job searchers' decisions, and continue searching. Individuals accept jobs in part by evaluating the expected income they will receive against their current income, as well as a range of other factors such as whether or not a spouse is working and the travel costs of commuting to a job.

Simulation Model

The simulation model of the career choices of biomedical graduates was implemented in NetLogo. Test data for the simulation were taken from the actual survey data from AAAS, which includes almost 9000 post-doctoral recipients or former postdoctoral recipients from 2004-2010. Additionally, data were taken from the AAAS survey of post-

² The dashed lines in the circle do not sub-divide the search area at all, they are simply used for graphical illustration. The search area itself can be understood as both a reflection of geography as well as individual preferences (e.g., family status, wages) that influence where individuals will search for jobs.

doctoral supervisors, which includes 1700 individuals over three survey years. This model will be rerun over time as we integrate more data from the SDR and the work completed by George Washington University.

Using this data we implement the following logical framework (as outlined in the pseudo-code as detailed in Figure 3.

Figure 3: Pseudo-Code for NetLogo Implementation

```

Initialize
    Set model parameters to zero
Create agents (job seekers and jobs) and allocate them randomly on the grid
Search jobs
    For each job seeker
        Calculate search distance
        For each job
            If job is in search distance and matches with job seeker's interest
then
        Calculate job seekers' knowledge and skills
        If |job seeker's knowledge and skills – job requirements| <= tolerance
level then Apply for the job
        end job
        end job seeker
Job offer and evaluate offer
    For maximum number of offers
        For each job
            Select the applicant with the highest knowledge and skills
        Make an offer
        end job
        For each job seeker with an offer
        For each offered jobs
            Calculate job's utility cost
            If job's utility cost < job seeker's current job
utility cost then
                Accept the offer
                if job seeker is PhD and job offered is postdoc job, then convert PhD to
postdoc
                if job seeker is postdoc and job accepted is postdoc, keep the job
seeker in the system
            else remove job seeker from the system
            Remove the job from the system
            end offered jobs
        end job seeker with an offer
    end maximum number of offers
Update agent numbers and characteristics
end t time periods

```

Next Steps

During the second year of the SWAM activities Ohio State and MIT will be working to develop the simulation model using the SDR data as supplied by the team from GWU. Currently, the groups have exchanged data requirements and the OSU/MIT team will revise the ABM model after receiving the data from GWU.